Cloud Economics

Cloud economics is a branch of knowledge concerned with the principles, costs and benefits of cloud computing.

Because IT leaders are constantly challenged to deliver information technology (IT) services with the greatest value for the business, they must determine specifically how cloud services will affect an IT budget and staffing needs. In assessing cloud economics, IT leaders weigh the costs pertaining to infrastructure, management, research and development (R&D), security and support to determine if moving to the cloud makes sense given their organization's specific circumstances.

Although the cloud can facilitate resource provisioning and flexible pricing, there are several cloud computing costs beyond instance price lists to consider. Pricing usually includes storage, networking, load balancing, security, redundancy, backup, software services and operating system (OS) licenses. IT leaders within an organization must closely examine the economics of moving to the cloud before deciding whether to invest in the expertise and time that is required to maximize cloud investments.

Therefore, the following questions need to be addressed before moving the applications to cloud:

- Is it always beneficial to go on cloud?
- When to go, when not to go?
- Whether one needs to buy something on cloud, or in-house infrastructure is suitable?
- How to balance between the two?

The following factors need to be considered from economic view point.

1) Common Infrastructure
    - Pooled, standardized resources, with benefits generated by statistical multiplexing.
2) Location independence
3) Online connectivity
    - Ensures service access
4) Utility pricing
    - Usage-sensitive or pay-per-use pricing, with benefits applying in environments with variable demand levels
5) On-demand resources
    - Scalable resources provisioned and de-provisioned without delay or costs associated with change.

With these considerations we look at it from valuation point.

Value of common infrastructure

1) **Economics of scale**
    - Reduced overhead costs
        - Example
            - For computer, overhead cost of AC, maintenance of UPS, AMC of the system, human resources etc. to maintain the system.
            - Buying the things is fine but it becomes costlier to maintain over a period of time.
            - Also the technology may become obsolete after some time.
    - Buyer power through volume purchasing
        - There may be certain rights or privileges being given to the buyer in case the product is purchased in bulk.
            - The rights may no longer be valid for installing new systems bought after some time.

2) **Statistics of scale**
    - For infrastructure built to peak requirements
        - Multiplexing demand may help in higher utilization.
    - For infrastructure built to less than peak
        - Multiplexing demand may help in reducing the unserved demand.
3) **Coefficient of variance (Cv)**
    - It is a statistical measure of dispersion of data points into in a data series around the mean.
    - Represents ratio of standard deviation to the mean (σ/μ)
    - Useful for comparing the degree of variation from one data series to another.
        - Example
            - In investment world the Coefficient of variance allows to determine how much volatility (risks) one is assuming in comparison to the amount of return expected from the investment
        - In simple language, the lower the ratio of standard deviation to mean return, the better the risk-return trade off.
    - In a way, it is "measure of smoothness", which implies that a facility with fixed assets servicing highly variable demand will achieve lower utilization than a similar one servicing relatively smooth demand.
    - **Multiplexing demand from multiple sources may reduce the coefficient of variation**.
        - **Adding n independent demand reduces the C_v by $1/\sqrt{n}$**

Economics is also related to type of workload.

- Peak workload will lead to congestion.
    - For example, if all classes break at 4 pm, congestion of traffic like cars, cycles, people etc will be more.

Value of location independence

We don't need to go to the actual hardware site. Applications, services and contents can now be accessed from wherever we are. It can be done through networks (wired, wireless) and satellites etc.

- Latency in computing
    - It is the delay before transfer of data begins following an instruction for its transfer.
        - Poor performance due to network latency
    - Human resource latency can be 10 sec to 100 millisecond.
    - Latency is correlated with:
        - Distance (strongly)
        - Routing algorithms of routers and switches.
    - Speed of light in fiber is only 124 miles/millisecond
    - Example:
        - If we are searching for a word in Google and it takes more than few seconds then we are not happy with that
        - If the VoIP (Voice over IP) has latency of 200 milliseconds or more then the transmission is not considered to be good.
            - Difficult to communicate
    - Supporting a global user base requires a dispersed service architecture (appropriate distributed architecture) which should provide coordination, consistency, availability and partition-tolerance.
        - It has direct implication on investment.

Value of utility pricing

As discussed earlier, economy of scale might not be very effective, but cloud services don't need to be cheaper to be economical.

Example: Consider a car

- Buy a car
- Rent a car for Rs 5000/- per day
- If the car is required for a couple of days for a trip, buying would be much more costlier than renting.
- if the car is needed daily for commuting then buying a car would be cheaper than renting.
- Therefore, it depends on demand.

D(t)	Demand for resource 0 < t < T
P	Max D(t): Peak demand
A	Avg ((D(t)): Average demand
B	Baseline (owned) unit cost
C	Cloud unit cost
U = C/B	Utility premium

Total cloud cost, $C_T = A * U * B * T$

Total Baseline cost, $B_T = P * B * T$

- Baseline should handle peak demand

When is cloud cheaper than owning?

- When $C_T < B_T$

  ⇨ $(A * U * B * T) < (P * B * T)$

  ⇨ $U < P/A$

This implies that the cloud is cheaper than owning when utility premium is less than ratio of peak demand to average demand.

Utility pricing in Real World

- In practice demands are often highly variable
  - For example, News, stories, marketing promotions, product launches, tax season, festival shopping etc.
- Often a hybrid model is best.
- Key factor is again the ratio of peak to average demand.
- But other costs should also be considered like network cost (both fixed cost and usage costs)
- Consider reliability and accessibility.

Value of On-Demand services

Problem faced: When owning your resources, you will pay a penalty whenever your resources do not match the instantaneous demands. In that case,

- Either pay for unused resources, or
- Suffer the penalty of missing service delivery

D(t): instantaneous demand at time t

R(t): resources at time t

Penalty cost $\propto \int_0^t [\, D(t) - R(t)] \, dt$

- If demand is flat, penalty = 0.
- If demand is linear, periodic provisioning is acceptable.
- If demand is non-linear (exponential): D(t) = e^t, any fixed provisioning interval (t_p) according to the current demands will fall exponentially behind.
- As Penalty grows exponentially, it becomes extremely difficult to manage.